Assessing Rural Transformations: Piloting a Qualitative Impact Protocol in Malawi and Ethiopia

James Copestake and Fiona Remnant

Centre for Development Studies, University of Bath

Bath Papers in International Development and Wellbeing

Working Paper No. 35

November 2014

The Centre for Development Studies at the University of Bath is an interdisciplinary collaborative research centre critically engaging with international development policy and practice.

Centre for Development Studies
University of Bath
Bath BA2 7AY
United Kingdom

http://www.bath.ac.uk/cds

Published by:
The Centre for Development Studies
University of Bath
Claverton Down
Bath, BA2 7AY, UK
http://www.bath.ac.uk/cds/

ISSN 2040-3151

Series Editors:
Susan Johnson
Althea-Maria Rivas

Assessing Rural Transformations: Piloting a Qualitative Impact Protocol in Malawi and Ethiopia

James Copestake and Fiona Remnant

Abstract

This paper contributes to the literature on qualitative approaches to impact evaluation, particularly in complex contexts. It reports on substantive and methodological findings from four pilot studies of a protocol for qualitative impact evaluation of NGO sponsored rural development projects in Malawi and Ethiopia. Two of the projects aimed to build resilience to climate change through support for a spectrum of livelihood diversification activities, while two focused on smallholder involvement in the value chains of specific cash crops. The protocol was designed and tested through action research with the aim of generating evidence in a credible, timely and cost-effective way to confirm the causal theories underpinning project actions, as well as to explore incidental sources of change and unanticipated effects. The paper describes the methodology, provides an overview of findings and reflects on lessons learnt in addressing problems of attribution, confirmation bias and generalizability. It suggests scope for further development of responses to these issues based on self-reported attribution, partial blinding of respondents and nesting qualitative evaluation in quantitative monitoring.

Key words: Impact evaluation, qualitative methods, food security, climate change adaptation, rural livelihoods, Malawi, Ethiopia, NGOs, confirmation bias, mixed methods, attribution

Acknowledgement

The authors are grateful to Moges Belay, Tefera Goshu, Peter Mvulu, Zenawi Zerihun, their field teams and interviewees for pioneering data collection using the QUIP. We are also grateful for help and support from staff of Self Help Africa and Farm Africa in Ethiopia, Ireland, Malawi and the UK, and to Myriam Volk who made a significant contribution to the project by helping to test the QUIP analysis methodology using NVivo.

The Excel engineers at F1F9 in Bath and New Delhi continue to provide generous pro-bono support to the project in the form of ever more ingenious spreadsheets, and we are very grateful for their time and ideas. Lastly, thanks to Laura Camfield, Keetie Roelen and two anonymous reviewers for comments on an earlier draft.

The work is supported by the Economic and Social Research Council/Department for International Development Joint Scheme for Research on International Development (Poverty Alleviation) grant number ES/J018090/1.

1 Introduction

This paper reports on pilot testing of a qualitative impact protocol - referred to as the QUIP - that aims to provide credible, timely and cost-effective evidence of impact based on the testimony of intended beneficiaries of rural livelihood interventions without the need for a control group. The QUIP aims to address the perennial question of how international development agencies evaluate the impact of their work, with particular reference to the challenges faced by NGOs seeking to assist smallholder farmers with often complex agricultural and rural livelihood transformations associated with market integration and adaptation to climate change. Evidence of programme impact is potentially useful both for organisational learning and for building legitimacy through improved external accountability. Its importance has been reinforced by the seemingly inexorable rise of results-based and performance management culture in development practice (Gulrajani, 2010; Ramalingam, 2013) notwithstanding concern that this approach is undemocratic (Eyben, 2013) and can encourage what Natsios (2010) refers to as "obsessive measurement disorder." While often framed in technical terms, the issue of how the impact of development interventions can realistically and credibly be evaluated has been one battleground for these debates (Camfield & Duvendack, 2014).

A central methodological issue is attribution: or how particular outcomes can reliably be causally linked to specific projects, interventions or mechanisms in different contexts. The dominant approach defines impact as the difference in the value of an outcome indicator (Y_1) for a given population after a particular intervention or 'treatment' (X) compared to what the value would have been for the same population if the treatment had not occurred (Y_0) (White, 2010:154). Putting aside the problem of consistent measurement of X and Y, a central issue is then how to establish a plausible counterfactual. If the evaluator can make a large number of observations of X and Y then they can draw on well-known quantitative approaches to address this problem, including the use of randomized control designs. In contrast, the research summarised in this paper addresses the scope for more qualitative and 'small n' approaches. Our motivation for this is that while there are a range of established qualitative impact evaluation methods to choose from (process tracing, for example) is the view that there has been insufficient empirical research into how best to employ and to adapt these to disparate kinds of development activities (Stern *et al.*, 2012:1; White and Phillips, 2012:5).

Among various criticisms of quantitative approaches that rely on experimental or quasi-experimental designs perhaps the most important concern is the feasibility of addressing the practical threats to internal validity.[1] In an immensely complex, diverse, fast changing, emergent and recursive social world many researchers have argued that it is simply too slow and expensive to generate sufficient data using experimental or quasi-experimental designs. . It may be possible to measure a large vector of variables **Y** for a given population and time period, and to

[1] Randomization is also no guarantee against pro-project bias (White, 2010:156), particularly if Y is obtained from respondents (and/or by researchers) who are not blind to whether they belong to the treatment or control sample, and may therefore be prone to different degrees of response bias, including Hawthorne and John Henry effects (Duvendack *et al.* 2011). For further discussion see Camfield and Duvendack (2014).

demonstrate how they are affected by exposure to a vector of interventions or treatments **X**. But each set of results is specific in time and space to a vector of confounding or contextual variables (**Z**) that is too small to be measured reliably, or too quickly becomes outdated in history (Pawson and Tilley, 1994). Realist evaluation offers one counterpoint to this, emphasising the need for a cumulative process of broadening understanding of context-mechanism-outcome interactions or knowledge of "…what works for whom in what circumstances, in what respects, over which duration… and why" (Pawson and Manzano-Santaella, 2012:177). This pursuit of realism can be viewed as being achieved at the expense of the precision gained from experimental methods, which generate statistically significant results through artificially restricting variation in treatment and contextual vectors (Levins, 1966). In this sense, the quest for alternatives to precise quantitative methods of impact evaluation entails dealing with "organised complexity" on its own terms, rather than through a process of deliberate reduction into a closed model with a more manageable number of variables and/or statistical properties.[2]

An alternative to estimating a counterfactual on the basis of statistical comparisons between respondents subject to different levels of exposure to a project/treatment is simply to ask intended beneficiaries what they think. If we are interested in finding out whether particular men, women or children are less hungry as a result of some action it seems ethically important as well as common-sense just to ask them (Anderson *et al.*, 2012). But even putting aside problems of construct validity (over the definition of hunger, for example) it is not obvious how easily they will be able to attribute changes in their experience to specific activities. And there may also be reasons to doubt the reliability of their responses, including *confirmation bias* (Haidt, 2012:93) or a tendency to anchor their responses to what is familiar or expected (Kahneman, 2011). In this paper I will also use the term *pro-project bias* to refer to the possibility that someone consciously or otherwise conceals or distorts what they think they know about an activity in the hope that doing so will reinforce the case for keeping it going. The instrumental value of asking people directly about attribution is practical and empirical. To what extent is it possible to find ways to benefit from their direct experience of the impact of a project in a way that is not undermined by potential pro-project bias?

The structure of the paper is as follows. The remainder of Section 1 elaborates on what is meant by a *credible* evaluation. Section 2 provides a short factual description of the methodology underpinning the QUIP, as designed and tested on two NGO projects in Malawi and two in Ethiopia. Section 3 presents selected findings from these pilot studies to illuminate the methodological discussion. Section 4 discusses three key methodological issues – attribution, confirmation bias and generalizability – and Section 5 concludes.

1.1 Defining credible impact evaluation

White (2010:154) notes that the term impact evaluation is widely used to refer both to any discussion of outcome and impact indicators, and more narrowly to studies that explicitly seek to attribute outcomes to a specified intervention. This paper adopts the second definition. It also

[2] Complexity is much discussed, but often rather loosely. For discussion of the term "organised complexity" see Ramalingam (2013:134). Here we take it to mean that the influence of X on Y is confounded by factors **Z** that are impossible fully to enumerate, of uncertain or highly variable value, difficult to separate, and/or impossible fully to control. Additional complexity arises if the nature and value of X and/or Y is also uncertain.

allows for the possibility that specific impact assessment methods (including those within apositivist tradition) can be nested within broader (including interpretive) evaluation approaches. Attributing impact is only one issue that evaluation addresses – others including how an intervention works, and whether it constitutes value for money (Stern et *al.*, 2012:36).

As a servant of action in a changing context the scientific rigour of impact evaluation also has to be weighed alongside cost, timeliness and fitness to purpose. Without rejecting the quest for consensus about what constitutes quality in qualitative research, Hammersley (e.g. 2013:83) also favours use of the term credibility rather than scientific rigour as a criterion for assessing impact evaluation, echoing the more general distinction between reasonableness and rationality (McGilchrist, 2010).[3] By credibility, I refer to one party being able to offer a sufficient combination of evidence and explanation to convince another party that a proposition is reasonable in the sense of being sufficiently plausible to act upon – not rational in a logical sense, perhaps, but neither irrational. While this emphasises the importance of context and trust, the rigour with which conclusions about impact are logically derived from stated evidence and assumptions is also clearly important.[4] A more specific approach to defining credibility with respect to impact evaluation is to agree on what constitutes reasonable evidence of causation. For example, an evaluator's claim to establishing impact (i.e. X causing Y in particular contexts) might be regarded as being credible if: (a) there is strong evidence that X and Y happened in such contexts, (b) X is described by a diverse range of stakeholders as having been a necessary component of a package of actions that are sufficient to cause Y in those contexts, (c) their explanations of the mechanism by which X caused Y in those contexts are independently arrived at and mutually consistent, (d) the counter-hypothesis that they have other reasons for making the statement can reasonably be refuted. The point is not to secure universal agreement, but to be as clear and precise as possible about what can reasonably be expected in a given context. For example, our emphasis here being on qualitative methods, the definition excludes the requirement for (e) evidence of how *much* Y varies according to exposure to X.[5]

[3] McGilchrist (2010) suggest humans are all capable of thinking in two distinct and complementary ways. The first more rational, depersonalized and certainty seeking abstracts and simplifies, producing narrower, more precise and focused models of the world. The second aims to be reasonable, concrete, less certain, contextual, person rather than idea oriented, emphasising difference rather than sameness, quantification over meaning). It is associated with open forms of attention and vigilance, alongside broader, contextualizing and holistic ways of thinking. Much of the time we employ both together, and this confers immense potential evolutionary advantages: to think narrowly (as forensic hunter-gatherer) and broadly (as agile evader of other hunters) at the same time, for example. But that does not rule out individuals having a stronger predisposition towards one way of thinking over the other. Rowson and McGilchrist (2013:30) make clear that this "horizontal" distinction is complementary but distinct from the "vertical" one between "fast" and "slow" thinking made by Kahneman (2011).

[4] A common way of further elaborating on the credibility of evidence is to distinguish between the validity of an approach, and the reliability of results arising from its application in a particular context. However, we agree with Lewis and Ritchie (2003:270) that this distinction is harder to sustain and therefore less useful for qualitative impact evaluation given that no study can ever be replicated in precisely the same time and setting in order to identify how far results are sensitive to implementation rather than design.

[5] Although scope for quantification will be explored through a second round of pilot studies making greater use of on-going monitoring (IHM) data.

The idea of credible causation, based on reasonableness, can be further elaborated by specifying minimum conditions for mitigating the risks of systematic bias. The definition above, for example, proposes structures and processes of evaluation that reduce the plausibility of complicity among different stakeholders. This falls short of scientific certainty, but in complex contexts it is often as much as we can hope for, particularly given the possibility that efforts to aim higher may be counterproductive in terms of cost, timeliness and policy relevance. In other words, I am not suggesting that this definition is universal or even widely accepted, rather that it is a realistic one in contexts where overcoming the attribution problem is particularly difficult.

2 Methodology

This section reports on action research comprising the design and testing of a qualitative impact protocol (QUIP).[6] Initial piloting was conducted with four projects sponsored by international NGOs: two in Malawi and two in Ethiopia. Details of them are set out in Table 1. Projects 1 and 3 concentrated their activities (**X**) on specific crops, while Projects 2 and 4 incorporated a broader spectrum of activities intended to promote livelihood diversification. However, all of them aimed to strengthen the livelihoods and food security of selected rural households, enabling the QUIP to be designed around a common set of impact indicators (**Y**) listed in the second column of the table. The context of all the projects can be described as one of organised complexity arising from the presence of interconnected, uncertain and hard-to-measure confounding factors (**Z**) affecting the casual links between **X** and **Y**. In both Malawi and Ethiopia these include climate change, commercialisation (Collier & Dercon, 2009; Future Agricultures, 2014), the activities of other NGOs working in the same area, and the evolution of public policy (e.g. Chirwa & Dorward, 2013; Abro et al. 2014) and social protection (Wedegebriel, 2013). In contrast to quantitative impact assessment methods, the QUIP sets out to generate differentiated empirical evidence of impact based on narrative causal statements of intended project beneficiaries without the requirement to interview a control group. Evidence of attribution is sought through respondents' own account of causal mechanisms linking **X** to **Y** alongside **Z,** rather than by relying on statistical inference based on variable exposure to **X**.

Table 1. Summary of pilot projects, impact indicators and confounding factors

Interventions (X)	Impact indicators (Y)	Confounding factors (Z)
Project 1. Groundnut production and marketing (Central Malawi) Project 2. Livelihood diversification (Northern Malawi) Project 3. Malt barley production and marketing (Southern Ethiopia) Project 4. Livelihood diversification (Northern Ethiopia)	Food production Cash income Food consumption Cash spending Quality of relationships Net asset accumulation Overall wellbeing	Weather Climate change Crop pests and diseases Livestock mortality Activities of other external organisations Market conditions Demographic changes Health shocks

[6] It covers work carried out between November 2012 and May 2014 as part of the three year 'ART Project' programme of research into "assessing rural transformations". This is in turn funded under a joint call of the UK Economic and Social Research Council (ESRC) and Department for International Development (DFID) for research into "measuring development".

Draft written guidelines for the QUIP were prepared for a methodology workshop held in June 2013 and attended by staff from the University of Bath, the University of Malawi, Self Help Africa, Farm Africa, Evidence for Development, Oxfam UK and Irish Aid.[7] Each section was subject to detailed discussion at the workshop, and further refined through field testing of the protocol with two NGO projects in Malawi in November 2013, and two in Ethiopia in May 2014. The guidelines cover commissioning of impact assessment, its relationship to other impact evaluation activities, sample selection, data collection methods, briefing and debriefing the field researchers, facilitating interviews, data analysis, quality assurance and use of findings.[8]

Data collection by two field researchers for each pilot study was intended to last ten days, comprising four days of household level interviews, one day of focus group discussions and five days of data transcription. For the initial pilot studies in Malawi eight households were interviewed, and four focus groups were carried out; in Ethiopia the number of households was increased to 16,while the focus groups remained the same (sufficient to cover groups of older and younger men and women). Sample sizes were dictated primarily by constraints on time and funding, with all data collection restricted to one or two villages only, selected purposively as reasonably typical of the project area.

The field researchers were independently contracted by the University of Bath, acting as lead evaluator. They set up interviews and focus group discussions without any contact with the selected NGO or project staff, or indeed knowledge of the project being analysed. In the absence of this information the research team entered the field with an introductory letter to relevant local officials and a list of individuals in selected villages from which randomly to draw the interview sample. They introduced themselves to respondents as independent researchers conducting a study of general changes in the rural livelihoods and food security of farmers in the selected area. The purpose of this 'blinding' procedure was primarily to reduce potential for pro-project bias on the part of respondents, and is discussed in Section 4. It also minimised diversion of NGO staff time and effort into impact evaluation.

The household interview schedule started by asking respondents about changes in household composition. It then worked through a series of discrete sections covering different impact domains, to explore how changes in food production and other sources of real and cash income relate to changes in spending, food consumption, asset accumulation, relationships and overall wellbeing. Each domain section starts with an open-ended generative question and finishes with one or more closed questions, as summarised in the Appendix. Optional probing questions (also shown) were also available to help the interviews sustain and deepen the conversation. A final section asked respondents to list organisations they interact with from outside their village, and

[7] This in turn drew upon a QUIP designed during the 1990s to meet the specific needs of microfinance organisations that also linked in-depth impact interviews with routine quantitative monitoring of borrower or 'client' level indicators (see Imp-Act, 2004).
[8] A draft copy of the QUIP is available at http://www.bath.ac.uk/cds/projects-activities/assessing-rural-transformations/index.html

to rank their importance, thereby providing them with an additional opportunity to volunteer information about the NGO being evaluated.[9]

The researchers recorded narrative data in the field on a paper pro-forma, subsequently copying it into an Excel spreadsheet with an identical layout. They then passed the data to staff at the University of Bath for analysis. Their task - having also been briefed about details of the project - was to identify and code cause-and-effect statements embedded in the data according to whether they (a) *explicitly* attributed impact to project activities, (b) made statements that were *implicitly* consistent with the project's theory of change, (c) referred to drivers of change that were *incidental* to project activities. These statements were also classified according to impact domains and whether respondents described effects as positive or negative.[10] A similar process was followed for analysis of the focus group data.

Findings were fed back to the NGO in the form of a ten page report for each project (to a standard format), accompanied with an annex setting out the coded cause-and-effect statements in full. The body of these reports comprised a series of tables with frequency counts of different kinds of narrative statement. Simple quantification of responses in this way was not intended to support any kind of statistical claim. Rather it provided an initial indication of the extent of congruence in responses across the sample. At the same time the project reports encouraged readers to draw on the coded narrative statements, which were provided as an appendix. These statements were organised thematically making them easier to read, whilst retaining the richness of the original data.

3 Findings

Although asked only after open-ended discussion, we start with answers to closed questions from household interviews, as these reflect respondents' own overall assessment of the direction of change in selected impact indicators.[11] The data from Malawi shown in Table 2 refers to perceived changes over the previous two years. For Project 1 (groundnut) the data indicates positive change in food production, cash income, cash spending and food consumption for all but two respondents. For Project 2 (climate adaptation) the picture is more mixed, with six out of eight respondents reporting falling food production and three of them also reporting negative changes with respect to the other indicators. The final column refers to net asset accumulation, and in the majority of the cases this follows the pattern of responses to the other questions: positive changes being associated with asset accumulation (7 cases), and negative changes being associated with assets sales (2 cases) possibly as a coping strategy. But the

[9] This echoes the more holistic area approach to assessing impact adopted by both the WIDE programme in Ethiopia (Bevan, 2013) and PADev in West Africa (Dietz *et al.*, 2013).

[10] The analysis of the first two Malawi pilots was conducted in parallel by two analysts, one using bespoke Excel software and the other the qualitative analysis package NVivo. This served a quality control function (e.g. leading to identification of spreadsheet errors), and also stimulating discussion and reflection on how to improve both coding and presentation of findings. The field research teams also provided feedback on the field work process and results.

[11] It is worth noting at this point that only eight households were interviewed in each area in Malawi, as compared to 16 in Ethiopia. The pilot will interview 24 individual respondents in the next round in Malawi, giving us valuable information on the relative advantages of different sample sizes.

number of mixed responses is also noteworthy, including three cases where the direction of net asset accumulation bucked the trend of changes in the other indicators. Open ended interviews offered various explanations, including negative health shocks and positive remittance flows, illustrating the complexity of household livelihood systems and the environment within which they operate.

Table 3 presents similar data on perceived changes (this time over the previous two years) for the Ethiopia projects. Project 3 (malt barley) reveals a consistent pattern of increasing or stable food production, cash income and food consumption, with most respondents also reporting improvements in cash spending (i.e. overall purchasing power), net asset accumulation and rising overall wellbeing. In contrast, Project 4 (livelihood diversification) reveals a more mixed picture of change. Positive responses outnumber negative for perception of food production, food consumption and overall wellbeing, but it is the opposite way round for changes in cash income and purchasing power. This illustrates a recurring theme in narrative interviews of retail price inflation eroding hard won improvements in real income.

In the case of Project 1,3 and 4 the selected NGO was picked out by respondents as the main organisation working with them from outside their village, although its precise identity was often confused by reference to the name of the project and/or local partners including local government extension workers. The institutional landscape was particularly confused in the case of Project 2, where the selected NGO was coordinating a project that also involved several other local agencies. Resolving these identity issues and establishing precisely who was doing what in which localities emerged as an important preliminary task to coding of the narrative data. Precisely how the selected NGOs are labelled by 'their' intended beneficiaries within the institutional landscape is itself potentially insightful, e.g. some government and NGO projects were confused.

Table 2. Responses to closed questions: Malawi projects

Gen	Age	FP	CY	CS	FC	AA	Gen	Age	FP	CY	CS	FC	AA
Project 1 (n=8)							Project 2 (n=8)						
F	61	=	+	-	-	+	F	58	-	+	+	+	+
F	31	+	+	+	+	+	F	39	-	-	-	-	+
M	49	+	+	+	+	+	M	33	-	+	+	+	+
F	22	+	+	+	+	+	M	23	-	-	-	-	+
F	31	-	-	-	=	-	M	54	-	-	-	-	-
F	22	+	+	+	+	-	F	43	-	+	+	+	=
M	26	+	+	+	+	+	M	32	+	+	+	+	+
M	43	+	+	+	+	+	M	42	+	+	+	+	+

Notes: FP - Food Production; CY – Cash income; CS – Cash Spending; FC – Food consumption; AA – net Asset accumulation. See last column of Table 2 for details of questions.

Table 4 shows the number and type of cause-and-effect statements extracted from the narrative data, juxtaposing it with closed question data already discussed. The first number indicates the number of household respondents making a statement of this kind, and the second the number

of focus groups. In so doing we move from evidence of perceived change to evidence of attribution classified according to whether respondents volunteered statements that *explicitly* mentioned the project as causal drivers, made statements that were *implicitly* consistent with the project's theory of change or *incidental* to it (the note to Table 4 provides more precise definition). The first point to note here is the frequency of explicit positive statements about impact relative to the complete absence of negative statements. The fewer explicit positive statements about Project 2 reflect at least in part the fact that project activities were less advanced in selected villages, and there was some confusion over the withdrawal of another NGO from the area. Reference to incidental negative drivers of change (many weather related) were also higher for Projects 2 and 4. In contrast, many respondents of the study of Project 3 volunteered statements about positive incidental drivers of change. Those relating to increased food production referred either to livestock rearing, vegetable (including potato) production or both, often linked to the work of government Development Agents (DAs). There were also numerous references to the benefits of government training in nutrition and gender relations, adding up to a consistent story of a community of farmers that were highly tuned into and responsive to progressive government outreach. Having avoided linking the field researchers to the NGO in order to reduce the risk of pro-project bias towards the project it is likely that these responses collectively reflect a tendency towards positive confirmation bias towards government activities.

The frequency counts presented in Table 4 do not convey the detail and diversity of information about causal processes in the narrative data. For example, an interesting finding about Project 1 was the mental accounting through which farmers linked income from groundnut production to the cost of fertilizer purchases for their main maize crop: a rise in income from groundnuts being discounted as of little importance if offset by the rising price of fertiliser, even though this probably would have happened anyway. This also illustrates how discrete drivers were often interestingly linked – e.g. positive endorsement of help with purchasing livestock, but hedged by reference to disease and mortality problems. To give another example, there was explicit and implicit support for the NGO project activities in Project 4 (particularly irrigated vegetable production), but such statements were often combined with reference to the magnitude of the incidental negative drivers, particularly lack of rainfall. For example, one of the focus groups of older men was reported as saying the following: *"As the agricultural land is so small and not suitable for crop production, many development agents have been advising farmers and providing training on how they can use their land for alternative sources of income. Because of the drought our income has recently been reduced. But still many farmers are struggling to make use of the limited water in the check dams and hand dug wells to produce crops and vegetables to earn some money."*

Table 3. Responses to closed questions: Ethiopia projects

G	A	FP	CY	CS	FC	AA	WB	G	A	FP	CY	CS	FC	AA	WB
Project 3 (n=16)								Project 4 (n=16)							
M	28	+	+	-	+	+	+	F	33	+	+	+	+	+	+
M	45	+	+	+	+	+	+	M	38	-	-	-	+	+	+
M	58	+	+	+	+	+	+	M	37	+	+	+	+	+	+
M	28	+	+	-	+	+	+	F	52	+	-	-	=	-	+
M	40	+	+	+	+	-	-	F	52	-	-	-	=	-	-
M	38	+	+	+	+	+	+	F	40	-	=	+	+	+	+
M	67	+	+	+	+	=	+	F	47	+	+	+	+	+	+
M	30	+	+	+	=	+	+	F	27	-	-	-	-	+	=
M	40	+	+	-	=	=	=	F	51	=	=	+	+	=	=
M	31	+	+	+	+	+	+	M	50	+	=	-	=	-	-
M	26	+	+	+	+	+	+	F	40	=	=	=	=	+	=
M	50	+	+	+	+	+	+	F	45	+	+	-	+	+	=
M	60	+	=	=	=	?	=	M	43	=	-	-	=	=	=
M	55	+	+	+	+	+	+	F	46	=	-	-	=	=	=
M	49	+	+	+	+	+	+	F	38	+	-	=	=	=	?
M	65	=	=	=	+	-	=	F	50	=	=	+	=	+	=

Notes: G – Gender; A – Age; FP - Food Production; CY – Cash income; CS – Cash Spending; FC – Food consumption; AA – net Asset Accumulation; WB - wellbeing. See last column of Table 2 for details of questions.

It is very easy to pull out narrative quotations such as this to support specific points, but thereby also to present the evidence in a biased or cosmetic way to support prior views. To counter this danger the data analysis for each project included a process of inductively grouping and then systematically tabulating drivers of change mentioned by at least two respondents (cf. Benini *et al.* 2013). The main drivers identified in this way are summarised in Tables 5 and 6. Data in brackets again indicates the frequency with which the driver was mentioned in both household interviews and focus groups. Asterisks indicate those drivers that explicitly or implicitly support or negate project theory. One unsurprising finding here is that the same drivers were mentioned repeatedly in relation to different impact indicators: the importance of advice from Development Agents in the case of Project 3 for example. This repetition is nevertheless important. For example, it is no surprise in the case of Project 1 that groundnut production was widely cited as improving food production, cash income and spending, and the same for new varieties of barley for Project 3. But it is significant that as crops grown primarily for sale these factors are also mentioned as positive drivers of food consumption.

Table 4. Frequency of causal statements and responses to closed questions compared from semi-structure interviews (first number) and focus groups (second number).

	POSITIVE RESPONSES					NEGATIVE RESPONSES				
	Expl	Impl	Inci	Unat	Closed	Expl	Impl	Inci	Unat	Closed
Project 1 (n=8)										
Food production	5,2	2,1	1,0	0,0	6	0,0	0,3	1,0	0,0	1
Cash income	4,4	4,0	2,0	0,0	7	0,0	1,2	2,0	0,0	1
Cash spending	4,4	1,0	1,0	0,0	6	0,0	0,1	2,0	0,0	2
Food consumption	3,1	1,0	1,0	0,0	6	0,0	0,2	1,0	0,0	2
Relationships	4,1	1,0	2,0	0,0	Na	0,0	2,2	2,0	0,0	na
Asset accumulation	2,2	0,1	2,0	0,0	6	0,0	1,1	2,0	0,0	2
Project 2 (n=8)										
Food production	1,0	2,0	0,0	0,0	2	0,0	1,0	4,4	0,0	6
Cash income	1,1	5,3	0,0	0,0	5	0,0	1,0	3,4	0,0	3
Cash spending	0,0	2,1	0,0	1,0	5	0,0	0,0	5,3	0,0	3
Food consumption	0,0	4,0	1,0	1,0	5	0,0	0,0	2,4	0,0	3
Relationships	0,0	0,3	3,3	1,0	Na	0,0	0,0	1,2	0,0	na
Asset accumulation	0,0	3,0	0,2	3,0	6	0,0	0,0	0,3	0,0	1

Notes: The first number in each cell refers to how many household interviews yielded such statements, and the second to how many focus groups did so. The four 'types of statement' were defined as: Expl = change explicitly attributed to the project or explicitly named project activities; Impl = change confirming or refuting the specific mechanism or theory of change by which the project aims to achieve impact, but with no explicit reference to the project or named project activities; Inci = change attributed to other forces incidental to (not related to) the activities included in the project's theory of change; Unat = change not attributed to any specific cause. Domains refer to sections of the interviewing and focus group schedules. Analysts classified statements as positive or negative according to the impact on respondents' wellbeing as expressed by respondents themselves; an option to classify responses as 'neutral' or unclear in its impact on the stated domain was also available, and used in the coded transcript to highlight where is was unclear, but not used in the analysis tables.

Table 5. Most widely cited positive and negative drivers of change, Malawi projects

Domain	Positive	Negative
Project 1: groundnut seed, Malawi (n=8,4)		
Food production	NGO support for groundnut crop (4,0)* NGO advice on making manure (2,2)* NGO advice on small-scale irrigation (2,1)*	Low sale price for crops (1,3)*
Cash Income	NGO support for groundnut crop (5,3)* NGO pass-on livestock programme (3,2)* NGO support for farming as a business (3,0)*	Low sale price for crops (2,3)*
Cash spending	NGO support for groundnut crop (5,3)* NGO support for farming as a business (3,0)* Village savings and loan groups (3,0)	Increased prices, including food (0,3)
Food consumption	NGO support for groundnut crop (2,1)*	Increased prices, including food (0,2)
Quality of relationships	NGO support for farming as a business (1,1)*	Economic hardship (0,2)
Net asset Accumulation	NGO support for groundnut crop (2,0)*	
Project 2: Climate change adaptation, Malawi (n=8,4)		
Food production	NGO livestock rotation programmes (2,3)* Training in conservation farming (2,0)*	Poor weather conditions (4,4) Livestock diseases (2,0)*
Cash Income	NGO village savings and loan groups (2,3)* NGO small-scale irrigation projects (2,1) *	Low sale prices for crops (2,1)
Cash spending		Poor weather conditions (4,3) Low sale prices for crops (0,2)
Food consumption	Training in nutrition (0,2)	Poor weather conditions (1,4)
Quality of relationships	NGO training in financial management (0,3) NGO village savings and loan groups (0,2)	
Net asset Accumulation		Poor weather conditions (0,3)

Table 6. Most widely cited positive and negative drivers of change, Ethiopia projects

Domain	Positive	Negative
Project 3: Malt barley seed, Ethiopia (n=16,4)		
Food production	Agricultural advice from DAs (16,4)* New varieties of barley from NGO (13,4)* Advice from DAs on livestock rearing (8,4)	
Cash Income	Agricultural advice from DAs (13,4)* New varieties of barley from NGO (11,3)* Advice from DAs on livestock rearing (9,4)	
Cash spending	Agricultural advice from DAs (5,4)* New varieties of barley from NGO (2,2)* DA training in financial management (2,0)	Increase in market prices of food, fertiliser and clothes (8,0) Increase in contributions to govt bodies (2,0)
Food consumption	Agricultural advice from DAs (8,2)* Advice on diet and nutrition from HEAs (5,4) New varieties of barley from NGO (3,2)*	
Quality of relationships	Kabele training in gender equality (10,4) DA training in working together (10,2) New varieties of barley from NGO (1,1)*	Increased work demands and competition between households (0,2)
Net asset Accumulation	New varieties of barley from NGO (0,3)*	
Project 4: Livelihood diversification, Ethiopia (n=16,4)		
Food production	Increased fruit & veg production (4,4)* Goat rearing (1,3)* Beekeeping (1,1)* Purchase of ox, camel or cow (2,0)	Snow in August (& shortage of rain) (6,1) Lack of water / drought (3,3) Problems maintaining livestock (2,0)* Decreased labour (2,0)
Cash Income	Increased fruit & veg production (4,3)* Goat rearing (3,1)*	Lack of water / drought (5,3) Snow in August (& shortage of rain) (4,0) Decreased labour (2,0)
Cash spending	Increased fruit & veg production (3,1)* Employment abroad (2,0)	Increased prices (5,0) Lack of water / drought (1,3) Price of fertiliser (2,0)
Food consumption	Increased fruit & veg production (5,4)* Cheaper vegetables (2,0)*	Lack of water / drought (3,1)
Quality of relationships	Sharing ideas and resources in new farming practices (4,2)*	Lack of voluntary community support (3,1)
Net asset Accumulation	Goat rearing (6,4)* Increased fruit & veg production (1,1)* Purchase of ox, camel or cow (2,0)	Problems maintaining livestock (4,0)*
Overall wellbeing	Increased fruit & veg production (3,3)* Beekeeping (2,0)* Improved health services (1,1)	Lack of voluntary community support (1,1)

4 Discussion

This section critically reflects on methodological issues encountered in designing and piloting the QUIP. These are grouped into three. The first section reviews the potential of the QUIP to generate internally valid evidence of project impact, subject to the premise that confirmation bias and related problems can be addressed. The second reflects on the strategy for mitigating confirmation bias, and the third reflects on questions of sampling bias, timing and external validity. The article concludes with a preliminary assessment of the overall credibility and cost-effectiveness of the approach taking into account all these considerations.

4.1 Attribution

One motivation behind the action research presented here was to explore scope for addressing the problem of impact attribution not only through statistical inference based on variation in exposure of a population to project interventions but also through self-reported attribution, in the form of narrative statements from intended beneficiaries, explaining what happened to them over a period of time compared to what *would have* happened to them in the absence of the activities being evaluated. To put it another way, the attribution strategy being explored relies on respondents being able and willing to imagine and to communicate statements about change relative to a hypothetical counterfactual of zero exposure to particular activities. It is certainly not rare for us to communicate contingent statements of this kind to each other: "if I hadn't been at the meeting I would not have got the job", for example. The tougher questions concern how much information such statements can reliably carry in different contexts, and how explicitly the contingent nature of the statement needs to be spelt out. For example, to say "I got the job because I went to the meeting" implies causation, but is rather more relaxed. I might still have got the job, if I had gone to some other meeting instead.

The four pilot studies certainly generated lots of cause-and-effect statements of the kind X caused Y. But even if accepted as unbiased and truthful their interpretation is not easy. One observation that can be made is that relatively few statements attempted to assess the magnitude of observed impact. The most precise statements referred to the effect of new varieties of barley seed on yields (Project 3), while others downplayed the impact of project activities relative to larger forces like climate (Project 4). In line with discussion of sampling issues below, the frequency with which certain statements about impact were made constitutes evidence of their credibility rather than magnitude or importance. Hence in most cases the magnitude of the impact per household remains unknown, and so in isolation the QUIP should therefore primarily be viewed as a method for contribution analysis rather than impact assessment.

One strategy for addressing this limitation is to use the QUIP in conjunction with more precise quantitative monitoring of changes in key variables.[12] In a second round of pilot studies ongoing monitoring surveys will be used to estimate the magnitude of changes in food security, with the QUIP providing complementary qualitative evidence from respondents of the main causes behind these changes. This can at the very least help to establish limits to the magnitude of change that might conceivably be attributed to an intervention. For example, if monitoring reveals at some future date that an indicator, Y_1, of household disposable income on average rose by 2% between baseline and a repeat survey, it will still be possible for the intervention to

[12] In the case of the selected projects the NGOs are monitoring the food security of intended beneficiary households using the individual household method (IHM) developed by the NGO Evidence for Development (EFD). This approach is based on a combination of participatory rapid rural appraisal, structured household interviewing and simulation using bespoke software. Field data is used to generate estimates of how the production, exchange and transfer entitlements (in cash and kind) of a sample of households compare with estimates of their food consumption needs based on standardised nutritional requirements and food conversion ratios. Adult equivalent entitlements for a cross section of households are then compared with a benchmark absolute poverty threshold and can be used to simulate the heterogeneous impact of price, output, income and other shocks, as well as the impact of project interventions.

have had an average impact of more than 2% because it might have offset the negative impact of a change in some confounding variable, Z_1, such as rainfall. However, claims of impact in excess of observed changes would also need to be substantiated by evidence that these confounding causal effects were indeed present. If sufficiently detailed then QUIP data on causal mechanisms can be combined with monitoring data on the relative magnitudes of key variables to construct models with which to simulate the impact on **Y** of different combinations of **X** and **Z**. Armed with such estimates it would then be possible to make cost-benefit calculations in order to compare the cost-effectiveness of selected projects relative to alternatives.[13] This illustrates one example of the potential for synergy between qualitative and quantitative methods in impact evaluation that is quite different from combinations where one is used to frame the other sequentially, or they are used in parallel to obtain more robust results through triangulation.

4.2 Confirmation Bias

If one criticism of impact evaluation based on self-reported attribution is that it generates weak evidence on the magnitude of change, another potentially even more damming argument is that it is particularly vulnerable to confirmation bias, whether based on a respondent's effective willingness to please or a more strategic calculation that exaggerating impact can contribute to continued or further project support. Nor is the risk of bias confined to respondents. Researchers can also accentuate the importance of project interventions by downplaying or remaining ignorant of other influences on respondents' lives, particularly given the dominance of performance management culture in development practice, prompting evaluations to focus narrowly on assessing progress towards stated project goals (Picciotto, 2014:35).[14] In contrast the QUIP approach aims tobe even-handed in eliciting evidence on the impact of treatment and potentially confounding variables.[15] It thereby also seeks to redraw the balance between "exploratory" and "confirmatory" approaches to impact evaluation (Copestake, 2014).

The QUIP pilots attempted a robust response to potential confirmation bias problems by recruiting independent field researchers in a way that meant they were unaware of the identity of the project being evaluated and the NGO implementing it. This emphasis on avoiding pro-project bias appears to be in tension with the argument for placing project theories of change at the heart of impact evaluation to facilitate formulation of clear and testable impact hypotheses (cf. Ton, 2012). However, the piloting of the QUIP demonstrated that this apparent tension can at least partly be resolved by separating the role of data collection from that of analysis. In other words, an exploratory data collection stage of the QUIP was nested within, but contractually

[13] Mueller et al. (2014) propose an alternative approach that entails using more specific questions to encourage respondents to quantify hypothetical counterfactuals.

[14] In the absence of scope for placebos and double blind interviewing then even quantitative impact evaluation methods that incorporate a 'control' groups are prone to this problem - in the form of Hawthorne and John Henry effects for example. However, these problems can to some extent be mitigated by ensuring interview questions focus on general changes experienced by respondents, thereby concealing project intentionality and minimising (though never eliminating) differences in the way interviews with 'treatment' and 'control' respondents are framed and structured.

[15] The repeated mention of the significant impact of the work of government agricultural experts in Project 3 is a good example of this – whilst not part of the NGO's project, the positive effects of both were inextricably intertwined, and it was important to note this relationship.

separated from a confirmatory analysis stage. One feature of this strategy was the involvement of another agency to serve as lead evaluator: recruiting and briefing the lead researchers, providing them with lists of potential respondents from project staff, and then carrying out the data analysis by cross-analysing the narrative data against information on the goals, activities and intended outcomes of the project. The good news from the pilots is that it demonstrated this process of 'blinding' is indeed feasible. Lead researchers remained unclear which projects they were specifically helping to evaluate, yet the protocol nevertheless succeeded in generating a substantial amount of useful data about their impact.

At the same time the piloting experience revealed at least four limitations of this approach to dealing with confirmation bias. First, removing the association between field workers and the implementing NGO left a vacuum in the minds of respondents that they presumably filled with other possibilities.[16] In all cases the field researchers identified themselves as being affiliated with national universities; and while this may not have eliminated pro-authority bias entirely it perhaps encouraged respondents to be more honest and hopeful. But in at least one case (Project 3) there seems at least the possibility that pro-NGO project bias was replaced by a generalised pro-government bias.

A second problem is the replicability of the model used for these pilot studies. The pool of suitably qualified researchers (combining knowledge of local languages with social research skills) is limited, and being part of a UK university sponsored research project helped to recruit some of the best, which may be more difficult for NGOs to replicate over the longer term. Although our collaborators readily understood and entered into the spirit of conducting the work blind, it could easily be misconstrued in other contexts as distrustful and is in any case hard to guarantee or sustain. Ultimately, blinding is perhaps less important than building up the pool of qualified social researchers with professional commitment to high research standards of independent evaluation and research ethics.

Third, while field researchers were left in the dark about the project this was not the case for the role of data analysts for whom knowledge of project theory was necessary in in order to code whether it was consistent or not with the empirical evidence collected. This raises the question of how far they too might have been prone to bias in coding and interpretation of the data. Distinguishing between explicit and implicit attribution, deciding how far multiple positive and negative cause-and-effect statements can be unbundled, and aggregation of these into groups were three of the analytical tasks that proved difficult to do in a completely mechanical and objective way. However, this point should not be overstated: the subjective space for using the written transcripts is much smaller than that faced by respondents and researchers in constructing those narratives, and in principal the analytical role is also more easily audited,

[16] Anthropologist Thayer Scudder once recounted being told categorically by a villager in Zambia that he must be from the government. When asked why he thought this, the villager replied "only three sorts of outsiders come here: government people, missionaries and traders. And if a missionary or a trader then you're the worst of either I've ever met." The world has of course moved on, but there is still something satisfyingly robust about the generalisation that outsiders in rural areas have either political, commercial or religious motives (see Levine, 1972:56-58).

particularly since the coded transcripts are attached in full to the report (enabling readers to take issue with coding if they so wish).

Fourth, not having being fully transparent with respondents about the purpose of interviews raises deeper ethical issues. In the case of the QUIP this did not involve an outright lie: the field researchers did indeed come from national universities and the research was indeed motivated by a broad interest in the lives and livelihoods of farmers in the selected areas. Having explained this it was made clear to respondents that their participation was entirely voluntary, and that their anonymity would be protected. It is also unlikely that concealing the identity of the NGO caused any harm. However, farmers were nevertheless deprived of information that might have prompted them to withdraw or to give voice to stronger views about the NGO, whether positive or negative. Thus there is an unavoidable ethical choice to be made between adherence to categorical principles (such as being as fully transparent as possible) and pragmatism about means (being economical with the truth) in pursuit of hopefully sufficiently important ends (more reliable and useful evaluation). While it may accurately reflect human psychology, a more contentious issue for some may nevertheless be the decision to base research methods on implicit distrust in what other people will say when presented with a fuller explanation of why the data is being collected.

These ethical issues cannot be fully posed in isolation from the wider political economy of any impact evaluation as a mechanism for accounting for the use of scarce resources, and in relation to the cost and ethics of methodological alternatives. For example, one motivation for the QUIP research was to investigate methods of impact evaluation that (a) give voice to respondents' own explanations of change rather than inferring this indirectly from often rather simple comparisons of their behaviour and (b) avoid assigning some people or villages, randomly or otherwise, into a control group that entails questioning them even when they are not benefiting directly and immediately from the project being evaluated.[17] More fundamentally still there is the issue of how to balance evaluation practices with different development ends, with QUIP falling somewhere between more extractive survey approaches and more participatory and democratic approaches.

Overall, confirmation bias may significantly undermine the credibility of qualitative impact evaluation, and the QUIP pilots suggest ways of addressing this. But doing so does not come without having to make compromises, and since the extent of such bias is itself very hard to evaluate or quantify it is not easy to assess how much importance should be paid to this problem in methodological design.[18]

4.3 Generalisability

The reflections above have focused on credibility of what QUIP findings reveal about the impact of each project on selected respondents, but not on how generalisable these findings are

[17] Such respondents can be compensated with money, lottery tickets or other token gifts, but this raises still more ethical dilemmas.

[18] It would be possible to test the blinding approach by randomly informing some respondents but not others of the identity of the NGO evaluated. However, the problem would remain of how to assess the extent to which results could be generalised to other contexts.

beyond the relatively small sample of project participants actually contacted and the time period - of two years or less - covered by the questions they were asked. This section first considers selection over project space, within communities and over time. It then reviews scope for generalization beyond project boundaries and time horizons.

For monitoring surveys that aim at precise estimation of the typical (hence overall) value of selected indicators subject to acceptable levels of statistical significance there is a relatively well understood science for sample selection. In contrast, qualitative research is designed primarily to identify not only the main causal mechanisms affecting key indicators but also unexpected outcomes; thus criteria and processes for sample selection are unavoidably less precise. In the case of the QUIP, the ideal scenario would have been to randomly select a sub-sample of all households covered by systematic monitoring surveys, and keep open the option to augment the size of an initially small sample until it becomes apparent that additional interviews are not generating sufficient additional evidence to justify the effort. A relatively higher level of duplication of responses can be observed, for example, across the sample of 16 household interviews conducted for Project 3, for example, than for Project 4.

The pilot studies were not able to draw samples this simply, not least because randomly selecting respondents across large and scattered project areas would have massively increased the cost of finding and reaching respondents. Consequently, selection proceeded in two stages, with an initial purposive selection of one or two villages, followed by random selection of households from within them. The issue of how representative the selected villages were of the wider project area is not one that can be addressed by the procedure described above (augmenting a random sample opportunistically) because of the relatively small numbers involved. In practice, purposeful selection relied on secondary data, and the number of villages selected was limited by the constraint to limit data collection to five days for two researchers. Best practice combined two steps: documenting key sources of variation between sub-areas within the project area (e.g. agro-climatic, including altitude, and proximity to markets); and inviting knowledgeable local stakeholders to sort villages into like groups on the basis of what they anticipate being the most important sources of variation in project performance. This at least can clarify how far villages selected for qualitative studies compare with others across the project, as well as the extent of within project contextual homogeneity. It quickly became apparent, for example, that farming systems across Project 2 were hugely diversified (with maize, rice, sorghum and cassava competing as staples). In contrast in Project 3 the farming system was relatively homogenous, with barley dominant at intermediate altitudes, and giving way to wheat and oats at the lower and upper margins respectively of the project area. An additional and underestimated source of factor that affected the QUIP pilot studies was variation in the nature and timing of project activities between villages. For example, in the case of Project 4, households were earmarked for one of five distinct livelihood diversification packages, and data collection was restricted to one of the two villages where they had all been introduced.

The challenge of minimising or at least clarifying the extent of geographical bias is complicated by the need to ensure adequate coverage of variation in project effects *within* villages and indeed within households. For example, projects may accentuate differences in access to

resources between households, and feed intra-household tensions over gender and age specific allocation of labour, cash and other resources. One way to address the second problem is through multiple interviews within each household to provide greater detail of information and gender sensitivity, but at the extra cost of doubling up on interviewers, and having to invest time in reconciling potentially inconsistent data. Separate second interviews within each household can also be difficult to arrange (due to absences for work, for example), and resolving differences in answers risks creating or accentuating tensions within the household. For these reasons QUIP interviews during the pilot stage were limited to one per household, starting with the primary respondent identified from project lists (e.g. almost entirely men in the case of Project 3), but without ruling out participation of other household members. At the same time the QUIP pilots augmented household data with exploratory gender and age-specific focus groups to explore whether replicating discussions within small peer groups rather than a household setting might elicit different data.[19] For example, we hypothesised that respondents might be more likely to complain about gendered effects arising from a shift to cash cropping outside their own household and without having to refer to it specifically. Focus groups did throw up some interesting contrasts: younger people often being more positive about change than the elderly, for example. But Table 4 does not reveal a consistent difference across the four studies in the ratio of positive to negative statements collected through household interviews and focus groups.

In addition to respondent recruitment at the extensive and intensive margin, complex issues arise with respect to timing and frequency of interviewing (Camfield & Roelen, 2012; Devereux *et al.*, 2012; Woolcock, 2009). With many project interventions linked to the farming cycle the minimum period for assessing change is a year, while at the other extreme it is optimistic to expect farmers to provide a detailed account of how different drivers of change interacted over more than a two year period. However, data over more than two years is clearly necessary to address the sustainability of post project impacts, implying that repeat studies are essential - particularly for projects such as the ones considered here that are profoundly influenced by longer-term fluctuations and trends in market activity, climate, demography and even culture.[20] A potential strength of qualitative assessment is that findings are separable and additive – i.e. each additional interview can independently add to understanding. Additional studies can also be organised relatively quickly over time and across space – e.g. in response to findings generated by routine monitoring of key indicators. They are also potentially valuable early in project design and implementation to challenge project assumptions (Lensink, 2014).[21]

[19] More specifically the QUIP guidelines were for four focus group discussions per study (for younger men, younger women, older men and older women), with a minimum of three people present in each and a maximum of eight. The guidelines suggest inviting participation from additional members of selected households (other than the lead respondent), augmented by encouraging them to bring along a friend (the idea being to encourage freer peer discussion of more sensitive topics). In practice selection of participants across the four studies was more *ad hoc*, with only 38 out of 96 belonging to selected households.

[20] With respect to culture, the studies of Projects 3 and 4 both raised the question of how projects were responding to (and perhaps influencing) a shift towards more individualistic and competitive relations between neighbouring households, including having less time to share coffee and being less likely to offer help to those in need.

[21] The QUIP carried out for Project 2 in this study demonstrates this; it was too early in the lifecycle of the intervention to provide much information about the effectiveness of the project, but it did provide useful

Overall, there are practical constraints to how far scope for generalization can be increased through better sampling methods without also taking into account the budget available for impact assessment in relation to the heterogeneity of activities and contexts within and between projects, and over time. The four projects reviewed here illustrate how bespoke design of projects around time and space bound technological and market opportunities are critical to supporting livelihood diversification and adaptation. Hence while building concurrent impact evaluation into larger-scale development programming can help, expanding the portfolio of assessment methods that can be used flexibly and iteratively is also important. The goal of the action research reported here is to develop a QUIP with a unit cost of less than £5,000 (the budget used in these pilots), that can be conducted from start to finish in a few weeks and can be scaled up and adapted to reflect changing project activities and conditions. A second round of pilot studies is planned for 2015, and there is clearly scope for further work both over time and in other contexts.

5 Conclusion

This paper has presented results from a first round of pilot testing of a qualitative impact assessment protocol tailored to provide independent feedback on how rural livelihood and climate adaptation project are affecting household level production, income and food security. First, it has suggested that it is possible to address problems of attribution and contribution using qualitative as well as quantitative methods by relying on narrative accounts of drivers of change collected directly from intended beneficiaries, particularly if this can be combined with quantitative estimates of changes in key indicators through model based simulation (not described here). Second, it has identified some scope for addressing pro-project or confirmation bias through the use of independent evaluators distanced from implementation. Third, it has pointed towards the importance of strengthening scalable methods of research that can be used adaptively, particularly in conjunction with routine monitoring of key indicators. Despite many years of effort to improve monitoring and evaluation of rural development considerable scope remains for improvement. While the focus of the action research reported here has been on rural livelihood transformations and their effect on relatively familiar and uncontroversial indicators of economic security, there is potential also to explore how the ideas and methods of qualitative assessment being tested relate to methods being utilised in other areas of intervention and with other indicators of wellbeing.

At a more general epistemological level this paper is unapologetic in promoting improvement in impact evaluation through systematic research and testing. At the same time it implicitly recognises that success hinges upon building trusted and sustained collaborative relationships that erode the frequently made but over-drawn distinction between research and practice. It also recognises the limitations of a positivist approach to improving development in the face of overwhelming contextual complexity and multiple stakeholder interests that spawn diverse and competing interpretations of what constitutes credible and useful evidence. More specifically, responses to problems of attribution, confirmation bias and generalizability have to be assessed against standards of construct, internal, external validity and reliability simultaneously. Likewise

information on what respondents saw as the most significant positive and negative forces affecting their livelihoods.

the messy details of design, data collection, analysis and use also have to be tackled together. Action research, such as that reported in this paper need not be premised on rational production of universal best solutions. Rather its purpose is to spur progress towards a range of more reasonable better practices, recognising that they will still be contested.

References

Akram-Lodhi, A. H. 1997 'The unitary model of the peasant household: an obituary? ', *Economic Issues* 2(1): 27-42

Abro, Z.A., Alemu, B.A., Hanjra, M.A. 2014 Policies for agricultural productivity growth and poverty reduction in rural Ethiopia. *World Development*, 59:461-474.

Anderson, M B, Brown, D, & Jean, I. 2012. *Time to listen: hearing people on the receiving end of international aid.* Cambridge MA: CDA Collaborative Learning Project.

Benini, A., Chowdhury, W.S., Khan, A.A., Bhattacharjee, R. 2013 Reflections on research processes in a development NGO: FIVDB's survey in 2013 of the change in household conditions and of the effect of livelihood trainings. A research note. Dacca: Friends in Village Development Bangladesh (FIVDB).

Bevan, P. 2013 *Researching social change and continuity: a complexity-informed study of twenty rural community cases in Ethiopia, 1994-2015.* Mokoro Ltd. Oxford.

Camfield, L. and Duvendack, M. 2014 Impact evaluation – are we 'off the gold standard'? *European Journal of Development Research*, 26(1):1-12.

Camfield, L. and Roelen, K. 2012 *Chronic poverty in rural Ethiopia through the lens of life histories*. Working Paper 399. Brighton: Institute of Development Studies.

Chirwa, E. and Dorward, A. 2013 *Agricultural input subsidies: the recent Malawi experience*. Oxford: Oxford University Press.

Collier, P. and Dercon, S. 2009 *African agriculture in fifty years: smallholders in a rapidly changing world?* Rome: FAO. Expert meeting on how to feed the world in 2050. ftp://ftp.fao.org/docrep/fao/012/ak983e/ak983e00.pdf.

Copestake, J. 2014 Credible impact evaluation in complex contexts: confirmatory and exploratory approaches. *Evaluation*, 20(4):412-27.

Devereux , S., Sabates-Wheeler, R. and Longhurst, R, eds. 2012 *Seasonality, rural livelihoods and development*. London, Earthscan from Routledge.

Dietz, T., Bymolt, R., Belemvire, A., van der Geest, K., de Groot, D., Millar, D., Zaal, F. 2013 *PaDev Guidebook: Participatory Assessment of Development*. Leiden: University of Amsterdam, Tamale University for Development Studies, Expertise pour le Developpement du Sahel, ICCO Alliance, Prisma, Woord en Daad.

Duvendack, M., Palmer-Jones, R. , Copestake, J., Hooper, L., Loke, Y., & Rao, N. 2011 *What is the evidence of the impact of microfinance on the well-being of poor people?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Eyben, R. 2013 *Uncovering the politics of 'evidence' and 'results'. A framing paper for development practictioners.* . Institute of Development Studies. Brighton: IDS. www.bigpushforward.net.

Future Agricultures 2014. *Supporting small farmers to commercialise*. CAADP Policy Brief 11. Brighton: Institute of Development Studies. http://www.future-agricultures.org/

Gulrajani, N. 2010 New vistas for development management: examining radical-reformist possibilities and potential. *Public administration and development, 30(*2), 136-148

Haidt, J. 2012 *The righteous mind: why good people are divided by politics and religion*. London: Penguin.

Hammersley, M. 2013 *The myth of research-based policy and practice*. Los Angeles: Sage.
Imp-Act 2004. QUIP: Understanding clients through in-depth interviews. Practice Note 2. Brighton: IDS. http://spmresourcecentre.net/iprc/assets/File/PN2_QUIP.pdf.

Kahneman, D. 2011 *Thinking, fast and slow*. London: Allen Lane.

Lensink, R. 2014 What can we learn from impact evaluations? *European Journal of Development Research*, 26(1): 12-17.

Levine, D. 1972 *Wax and gold: tradition and innovation in Ethiopian culture*. Chicago: Chicago University Press.

Levins, R. 1966 The strategy of model building in population biology. *American Scientist, 54*(4), 11.

Lewis, J, & Ritchie, J. 2003 Generalising from qualitative research: a guide for social science students and researchers. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice* (pp. 263-286). London, Thousand Oaks, New Delhi: Sage.

Mayne, J. 2012 Contribution analysis: coming of age? *Evaluation*, 18(3):270-280.

McGilchrist, I. 2010 *The master and his emissary: the divided brain and the making of the Western World.* New Haven: Yale University Press.

Mueller, C.E., Gaus, H., Rech, J. 2014 The counterfactual self-estimation of programme participants: impact assessment without control groups or pretests. *American Journal of Evaluation*, 35(1):8-25.

Natsios, A. 2010 *The clash of the counter-bureaucracy and development*. Working paper Washington D.C: Center for Global Development.

Pawson, R., & Manzano-Santaella, A. 2012 A realist diagnostic workshop. *Evaluation, 18*(2), 176-191.

Pawson, R., & Tilley, N. 1994 What works in evaluation research? *British Journal of Criminology, 34*(3), 15.

Picciotto, R. 2014 Is impact evaluation evaluation? *European Journal of Development Research*, 26(1): 31-38.

Ramalingam, B. 2013 *Aid on the Edge of Chaos: Rethinking International Cooperation in a Complex World*. Oxford: Oxford University Press.

Rowson, J., McGilchrist, I. 2013 *Divided brain, divided world: why the best part of us struggles to be heard*. London: RSA Action and Research Centre. www.thersa.org.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. 2012 *Broadening the range of designs and methods for impact evaluations.* London: DFID.

Ton, G. 2012 The mixing of methods: a three-step process for improving the rigour in impact evaluations. *Evaluation, 18*(1), 20.

Wedegebriel, Z.B. and Prowse, M. 2013 Climate change adaptation in Ethiopia: to what extent does social protection influence livelihood diversification? *Development Policy Review*, 31(S2):35-56.

White, H. 2010 A contribution to current debates in impact evaluation. *Evaluation 16*(2), 11.

White, H., & Phillips, D. 2012 A*ddressing attribution of cause and effect in 'small n' impact evaluations: towards an integrated framework*. London: International Initiative for Impact Evaluation

Woolcock, M. 2009 *Towards a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy*. Working Papers. University of Manchester: Brooks World Poverty Institute.

The Centre for Development Studies (CDS), University of Bath

The Centre for Development Studies aims to contribute to combating global poverty and inequality through primary research into the practical realities of global poverty; and, critical engagement with development practice and policy making. In December 2011, the Bath Papers in International Development (BPD) working paper series was merged with the Wellbeing in Developing Countries (WeD) Working Paper Series, which has now been discontinued. The new series, Bath Papers in International Development and Well-Being continues the numbering of the BPD series.

Bath Papers in International Development and Well-Being (BPIDW)

Bath Papers in International Development and Well-Being publishes research and policy analysis by scholars and development practitioners in the CDS and its wider network. Submissions to the series are encouraged; submissions should be directed to the Series Editor, and will be subject to a blind peer review process prior to acceptance.

Series Editors: Susan Johnson and Althea-Maria Rivas

Website: http://www.bath.ac.uk/cds/publications

Email: s.z.johnson@bath.ac.uk or a.rivas@bath.ac.uk

2014

No. 34. "We don't have this is mine and this is his": Managing money and the character of conjugality in Kenya
Author(s): *Susan Johnson, Centre for Development Studies, University of Bath*

No. 33. Can civil society be free of the natural state? Applying North to Bangladesh
Author(s): *Geof Wood, Centre for Development Studies, University of Bath*

No. 32. Creating more just cities: The right to the city and the capability approach combined
Author(s): *Séverine Deneulin, Centre for Development Studies, University of Bath*

No.31. Engaging with children living amidst political violence: Towards an integrated approach to protection
Author(s): *Jason Hart, Centre for Development Studies, University of Bath*

No. 30. Competing visions of financial inclusion in Kenya: The rift revealed by mobile money transfer
Author(s): *Susan Johnson, Centre for Development Studies, University of Bath*

No. 29. Can't buy me happiness: How voluntary simplicity contributes to subjective wellbeing
Author(s): *Nadine van Dijk, United Nations Research Institute for Social Development, Switzerland*

2013

No. 28. Challenge funds in international development
Author(s): *Anne-Marie O'Riordan, James Copestake, Juliette Seibold & David Smith, Triple line Consulting and University of Bath*

No. 27. From the Idea of Justice to the Idea of Injustice: Mixing the Ideal, Non-ideal and Dynamic Conceptions of Injustice
Author(s): *Oscar Garza, Centre for Development Studies, University of Bath*

No. 26. Understanding Policy and Programming on Sex-Selection in Tamil Nadu: Ethnographic and Sociological Reflections
Author(s): *Shahid Perwez, Centre for Development Studies, University of Bath*

No. 25. Beyond the grumpy rich man and the happy peasant: Subjective perspectives on wellbeing and food security in rural India
Author(s): *Sarah C. White, Centre for Development Studies, University of Bath*

No. 24. Behind the aid brand: Distinguishing between development finance and assistance
Author(s): *James Copestake, Centre for Development Studies, University of Bath*

No. 23. The political economy of financial inclusion: Tailoring policy to fit amid the tensions of market development
Author(s): *Susan Johnson, Centre for Development Studies, University of Bath; and Richard Williams, Oxford Policy Management, Oxford*

No. 22. 'Everything is Politics': Understanding the political dimensions of NGO legitimacy in conflict-affected and transitional contexts
Author(s): *Oliver Walton, Centre for Development Studies, University of Bath*

No. 21. Informality and Corruption
Author(s): *Ajit Mishra, University of Bath; and Ranjan Ray, Monash University, Australia*

No. 20. The speed of the snail: The Zapatistas' autonomy *de facto* and the Mexican State
Author(s): *Ana C. Dinerstein, Centre for Development Studies, University of Bath*

No. 19. Patriarchal investments: Marriage, dowry and economic change in rural Bangladesh
Author(s): *Sarah C White, Centre for Development Studies, University of Bath*

2012

No. 18. Political economy analysis, aid effectiveness and the art of development management
Author(s): *James Copestake and Richard Williams, Centre for Development Studies, University of Bath*

No. 17. Justice and deliberation about the good life: The contribution of Latin American *buen vivir* social movements to the idea of justice
Author(s): *Séverine Deneulin, Centre for Development Studies, University of Bath*

No. 16. Limits of participatory democracy: Social movements and the displacement of disagreement in South America*; and,*
Author(s): *Juan Pablo Ferrero, Department of Social and Policy Sciences, University of Bath*

No. 15. Human rights trade-offs in a context of systemic unfreedom: The case of the smelter town of La Oroya, Peru
Author(s): *Areli Valencia, University of Victoria, Canada*

No. 14. Inclusive financial markets: Is transformation under way in Kenya?
Author(s): *Susan Johnson, Centre for Development Studies, University of Bath; and, Steven Arnold, Department of Economics, University of Bath*

No. 13. Beyond subjective well-being: A critical review of the Stiglitz Report approach to subjective perspectives on quality of life
Author(s): *Sarah C. White, Centre for Development Studies, University of Bath, Stanley O. Gaines, Department of Psychology, Brunel University; and, Shreya Jha, Centre for Development Studies, University of Bath*

2011

No. 12. The role of social resources in securing life and livelihood in rural Afghanistan
Author(s): *Paula Kantor, International Centre for Research on Women; and, Adam Pain, Afghanistan Research and Evaluation Unit*

2010

No. 11. Côte d'Ivoire's elusive quest for peace
Author(s): *Arnim Langer, Centre for Peace Research and Strategic Studies, University of Leuven*

No. 10. Does modernity still matter? Evaluating the concept of multiple modernities and its alternatives
Author(s): *Elsje Fourie, University of Trento*

No. 9. The political economy of secessionism: Inequality, identity and the state
Author(s): *Graham K. Brown, Centre for Development Studies, University of Bath*

No. 8. Hope movements: Social movements in the pursuit of development
Author(s): *Séverine Deneulin, Centre for Development Studies, University of Bath; and, Ana C. Dinerstein, Centre for Development Studies, University of Bath*

No. 7. The role of informal groups in financial markets: Evidence from Kenya
Author(s): *Susan Johnson, Centre for Development Studies, University of Bath, Markku Malkamäki, Decentralised Financial Services Project, Kenya; and, Max Niño-Zarazua, Independent Consultant, Mexico City*

2009

No. 6. ‘Get to the bridge and I will help you cross’: Merit, personal connections, and money as routes to success in Nigerian higher education
Author(s): *Chris Willott, Centre for Development Studies, University of Bath*

No. 5. The politics of financial policy making in a developing country: The Financial Institutions Act in Thailand
Author(s): *Arissara Painmanakul, Centre for Development Studies, University of Bath*

No. 4. Contesting the boundaries of religion in social mobilization
Graham K. Brown, Centre for Development Studies, University of Bath,
Author(s): *Séverine Deneulin, Centre for Development Studies, University of Bath; and, Joseph Devine, Centre for Development Studies, University of Bath*

No. 3. Legible pluralism: The politics of ethnic and religious identification in Malaysia
Author(s): *Graham K. Brown, Centre for Development Studies, University of Bath*

No. 2. Financial inclusion, vulnerability, and mental models: From physical access to effective use of financial services in a low-income area of Mexico City
Author(s): *Max Niño-Zarazua, Independent Consultant, Mexico City; and, James G. Copestake, Centre for Development Studies, University of Bath*

No. 1. Financial access and exclusion in Kenya and Uganda
Author(s): *Susan Johnson, Centre for Development Studies, University of Bath; and, Max Niño-Zarazua, Independent Consultant, Mexico City*